Team:Data Minds

Unpacking the Spotify Dataset: Data Analysis and Strategic Insights

This presentation details the full end-to-end data science process applied to a rich Spotify dataset, transforming raw track data into actionable strategic recommendations. We will cover everything from initial data loading and cleaning to advanced multivariate analysis and final synthesis.

Presented by
Nilanjana Saren|Anshika Pandey|Ekadashi Sarder|Dipanjan Halder

Structure:

1. Setup and Data Acquisition
2. Initial Assessment: Understanding Data Quality and Structure
3. Exploratory Data Analysis: Univariate Distributions
4. Bivariate Analysis: Feature Relationships
5. Multivariate Analysis: Interacting Forces
6. Time Series Analysis: The Evolution of Music
7. Outlier Detection: Identifying Anomalous Tracks
8. Final Insights and Strategic Recommendations

Chapter 1: Setup and Data Acquisition

1. Importing Required Libraries

We load essential Python libraries (pandas, numpy, matplotlib, seaborn) for robust data handling, complex numerical computation, and high-fidelity statistical visualization.

2. Data Importation

The Spotify track data CSV files are imported into the environment. We ensure data integrity by verifying successful loading and preliminary record counts for consistency.

3. Initial Dataset Overview

We perform a critical initial assessment of the data structure, utilizing methods like .info() to check types and missing values, .describe() for statistical summaries, and .shape for dimensionality.

Initial Assessment: Understanding Data Quality and Structure

```
tracks.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62317 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62317 non-null  object
 1   track_name        62317 non-null  object
 2   artist_name       62317 non-null  object
 3   year              62317 non-null  int64
 4   popularity        62317 non-null  int64
 5   artwork_url       62317 non-null  object
 6   album_name        62317 non-null  object
 7   acousticness      62317 non-null  float64
 8   danceability      62317 non-null  float64
 9   duration_ms       62317 non-null  float64
 10  energy            62317 non-null  float64
 11  instrumentalness  62317 non-null  float64
 12  key               62317 non-null  float64
 13  liveness          62317 non-null  float64
 14  loudness          62317 non-null  float64
 15  mode              62317 non-null  float64
 16  speechiness       62317 non-null  float64
 17  tempo             62317 non-null  float64
 18  time_signature    62317 non-null  float64
 19  valence           62317 non-null  float64
 20  track_url         62317 non-null  object
 21  language          62317 non-null  object
dtypes: float64(13), int64(2), object(7)
memory usage: 10.5+ MB
```

- The dataset contains details of 62,317 Spotify tracks with 22 fully populated columns covering track info, audio features, and popularity.image.jpg
- All numerical audio features use float data types, while track, artist, and album info are stored as text

```
desc
```

	Column Name	Description
0	track_id	A unique identifier for the track on Spotify.
1	track_name	The title of the song.
2	artist_name	The name of the artist(s) who performed the song.
3	year	The release year of the song.
4	popularity	A measure of how popular a track is, ranging f...
5	artwork_url	A URL pointing to the album artwork for the tr...
6	album_name	The name of the album the track belongs to.
7	acousticness	A confidence measure indicating whether the tr...
8	danceability	A measure of how suitable a track is for danci...
9	duration_ms	The duration of the track in milliseconds.
10	energy	A perceptual measure of intensity and activity...
11	instrumentalness	Predicts whether a track contains no vocal con...
12	key	The key the track is in, represented as an int...
13	liveness	Detects the presence of an audience in the rec...
14	loudness	The overall loudness of a track in decibels (dB).
15	mode	Indicates the modality (major or minor) of a t...
16	speechiness	A measure detecting the presence of spoken wor...
17	tempo	The overall estimated tempo of a track in beat...
18	time_signature	An estimated overall time signature of a track
19	valence	A measure from -1.0 to 1.0 describing the musi...
20	track_url	A URL to the Spotify track.
21	language	The detected language of the song's lyrics.

- The Spotify dataset contains 62,317 tracks and 22 columns, with each column describing either track info, popularity, or audio features like danceability and energy.image.jpg
- Column descriptions explain song characteristics such as tempo, loudness, instrumentalness, and language, enabling diverse music analysis

```
desc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22 entries, 0 to 21
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Column Name  22 non-null     object
 1   Description  22 non-null     object
dtypes: object(2)
memory usage: 484.0+ bytes
```

- The data dictionary contains 22 entries, each representing a unique column from the main dataset.image.jpg
- Both "Column Name" and "Description" fields are fully populated with no missing data.image.jpg
- Each column is stored as an object type, suitable for textual metadata and explanations.image.jpg
- The compact structure makes it easy to reference column roles and understand dataset organization.

```
desc.describe(include='all')
```

	Column Name	Description
count	22	22
unique	22	22
top	track_id	A unique identifier for the track on Spotify.
freq	1	1

- Each of the 22 columns in the dictionary is uniquely described, confirming no duplicates in either attribute names or descriptions.image.jpg
- The most common column, "track_id", and its description both appear only once, indicating all entries are distinct.image.jpg
- The descriptions provide clear explanations for data fields, supporting easy understanding and accurate data usage.image.jpg
- The dataset structure is fully populated and suitable for quick reference or documentation purposes

Exploratory Data Analysis: Univariate Distributions

Analyzing the distribution of individual features provides foundational insights into music trends and acoustic characteristics within the dataset.

Numerical Features

Histograms and Kernel Density Estimates (KDE) for metrics like **Energy**, **Danceability**, and **Valence** showed generally normal distributions, centering around moderate values. **Loudness** exhibited a left-skew, indicating most tracks are produced at high volume levels.

Categorical Features: Key Signatures

Count plots revealed that the key signatures of C Major (Key 0) and G Major (Key 7) are the most dominant across the dataset, suggesting a preference for these keys in popular music production.

Popularity

A measure of how popular a track is, ranging from 0 to 100

Duration

The duration of the track (here, in milliseconds)

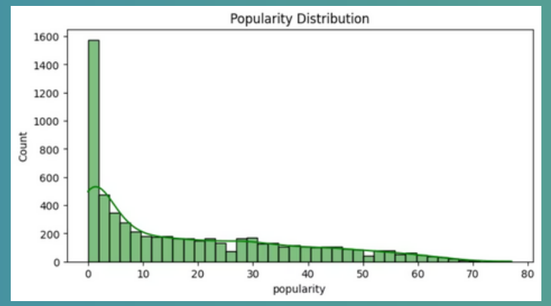
Tempo

The overall estimated tempo of a track in beats per minute(BPM)

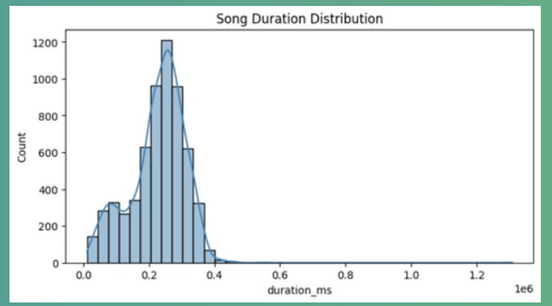
Danceability

A measure of how suitable a track for dancing ranging from -1.0 to +1.0

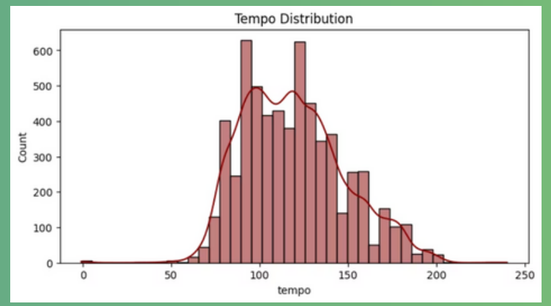
Numerical Features:-



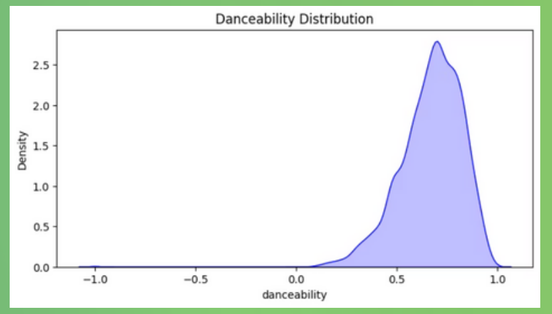
It shows a very high number of tracks with extremely low popularity scores, indicating that the majority of songs receive very little attention or engagement compared to others. This suggests that most tracks on the platform struggle to gain widespread recognition.



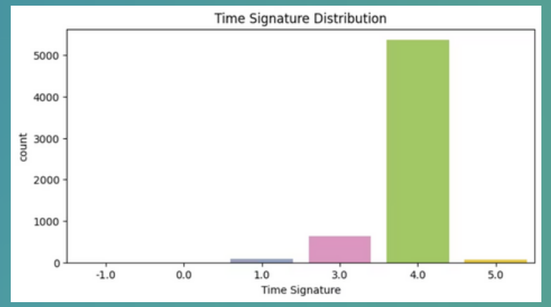
The highlighted region shows most song durations cluster between 150,000 and 300,000 milliseconds, with a peak around 220,000 ms, indicating typical songs last about 2.5 to 5 minutes and this duration range is the most common among tracks in the dataset.



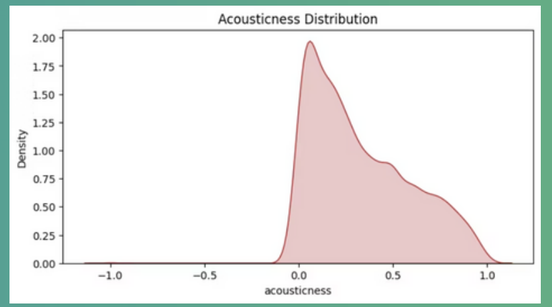
Most of the bars cluster around 100–130 BPM with a single main peak near roughly 120 BPM, indicating a unimodal, slightly right-tailed tempo distribution where mid-tempo songs are most common and higher-tempo outliers extend the right tail modestly.



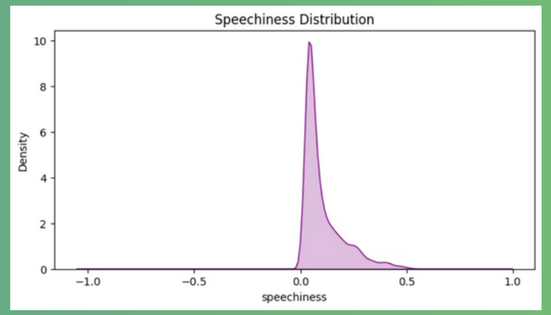
The density peak sits around danceability 0.6–0.8, showing most tracks are moderately to highly danceable, with few low-danceability songs and a slight right skew toward very danceable tracks.



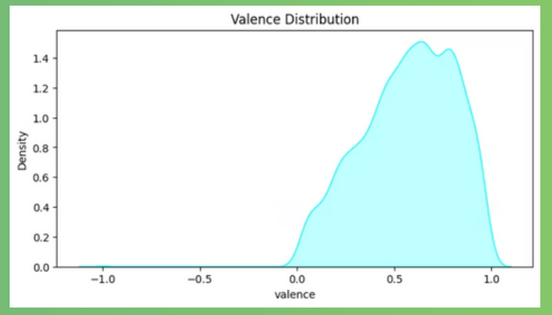
Most songs have moderate-to-high danceability, clustering between 0.6 and 0.8, indicating that energetic and dance-friendly tracks are common in the dataset.



Acousticness is concentrated at low-to-mid values (around 0.1–0.5), suggesting most tracks are primarily electronic or amplified rather than purely acoustic, with relatively few highly acoustic songs.

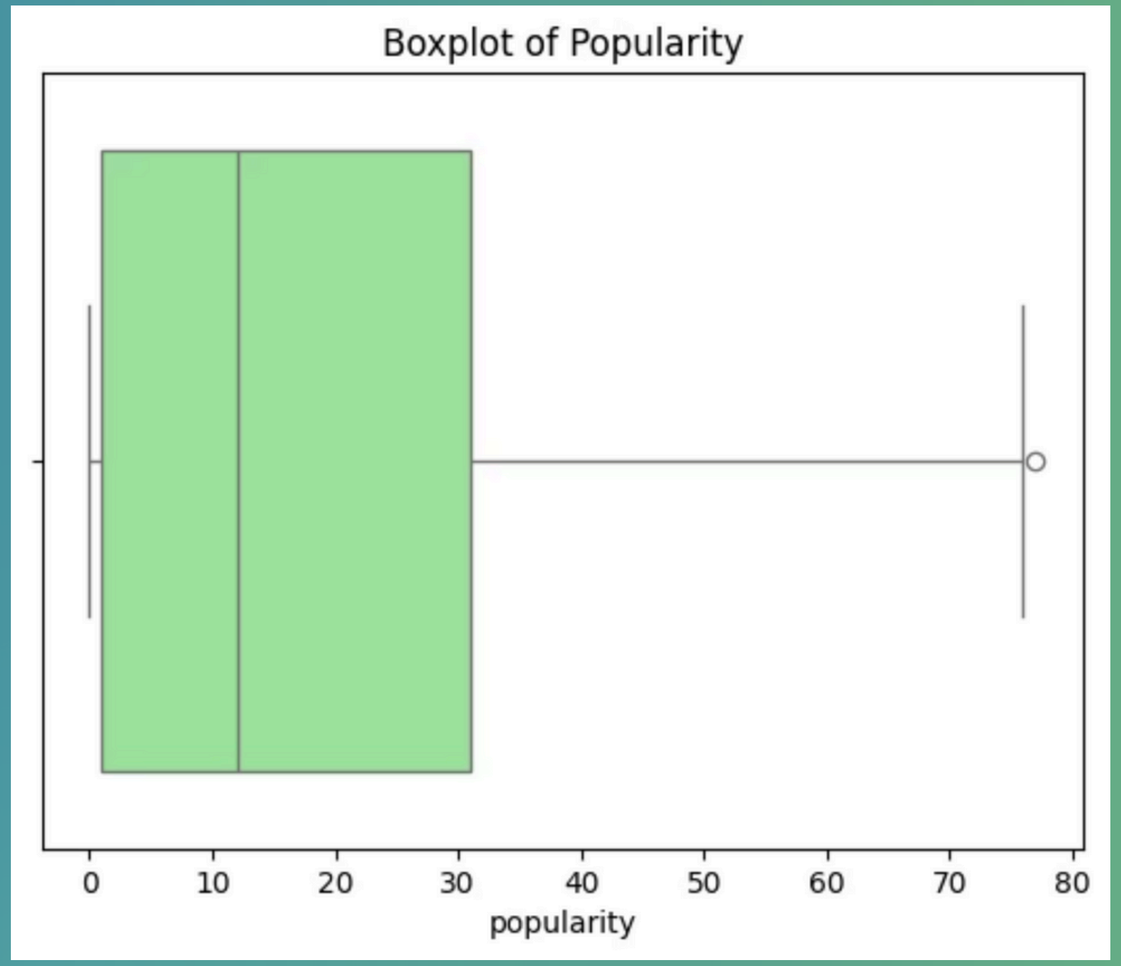


Speechiness is highly concentrated near zero, meaning most tracks contain very little spoken word content and are primarily sung or instrumental.

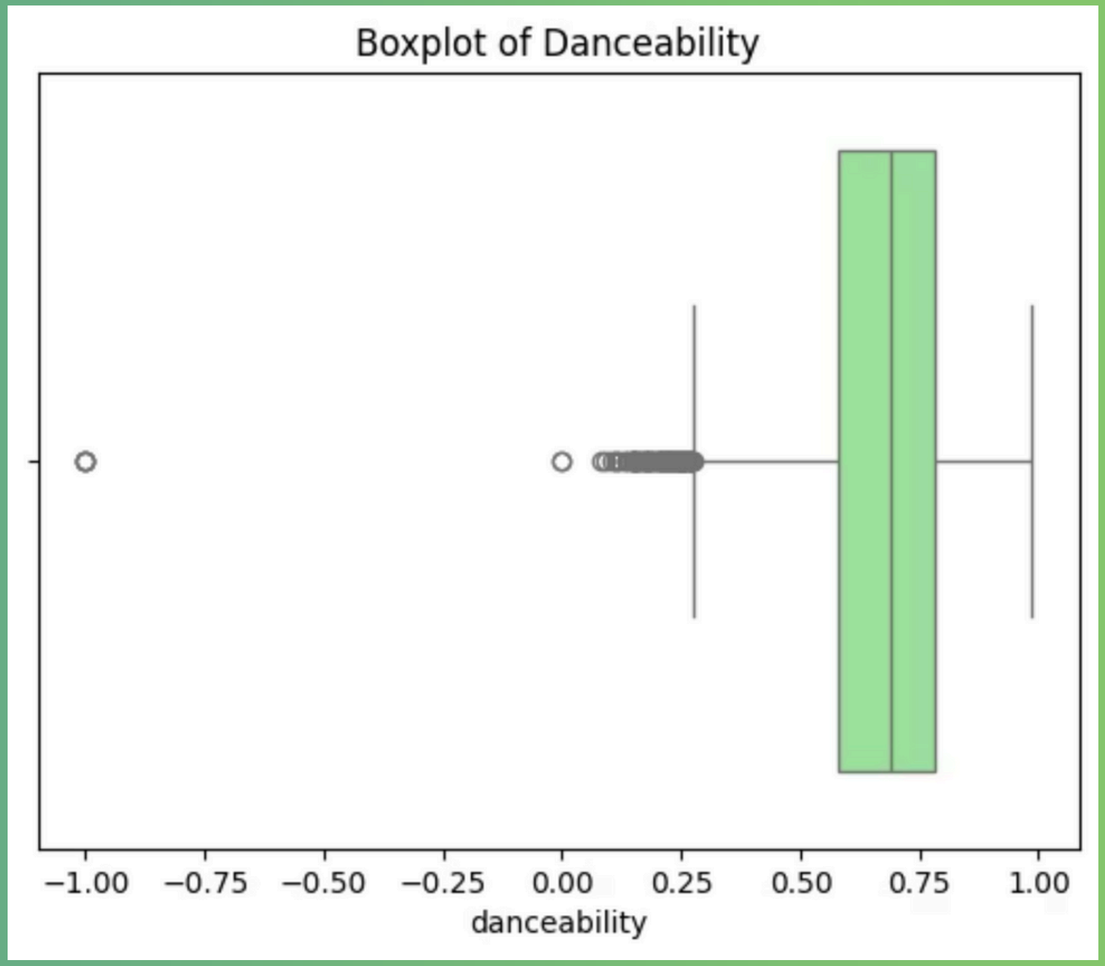


Valence skews toward 0.4–0.8, indicating the dataset leans positive and upbeat, with few very low or very high extremes; most tracks convey moderately happy moods suitable for mainstream listening.

Boxplot Insights: Popularity & Danceability of Songs



Most songs have low popularity scores, with the central range spanning 5–30 and a few tracks reaching much higher values as outliers.



Most tracks have high danceability with a median around 0.7, a tight IQR roughly 0.6–0.8, and a few low outliers indicating rare non-danceable songs.

Categorical Features: Key Signatures

For categorical univariate analysis, the most meaningful columns are:

- artist_name → Distribution of tracks by artist (top 10/20 artists).
- album_name → Distribution of tracks by album (top albums).
- language → Distribution of tracks by language

	Artist	Track Count	Percent of Dataset
0	Anirudh Ravichander	417	6.78
1	Yuvan Shankar Raja	384	6.24
2	Santhosh Narayanan	199	3.24
3	Devi Sri Prasad	151	2.46
4	Thaman S	140	2.28
5	G. V. Prakash	125	2.03
6	Hiphop Tamizha	117	1.90
7	Vijay Antony	50	0.81
8	Anirudh Ravichander, Dhanush	49	0.80
9	Harris Jayaraj	46	0.75
10	G. V. Prakash, Saindhavi	40	0.65
11	Harris Jayaraj, Karthik	37	0.60
12	Yuvan Shankar Raja, Premgi Amaren	35	0.57
13	Anirudh Ravichander, Jonita Gandhi	20	0.33
14	Thaman S, K. Pranati	20	0.33

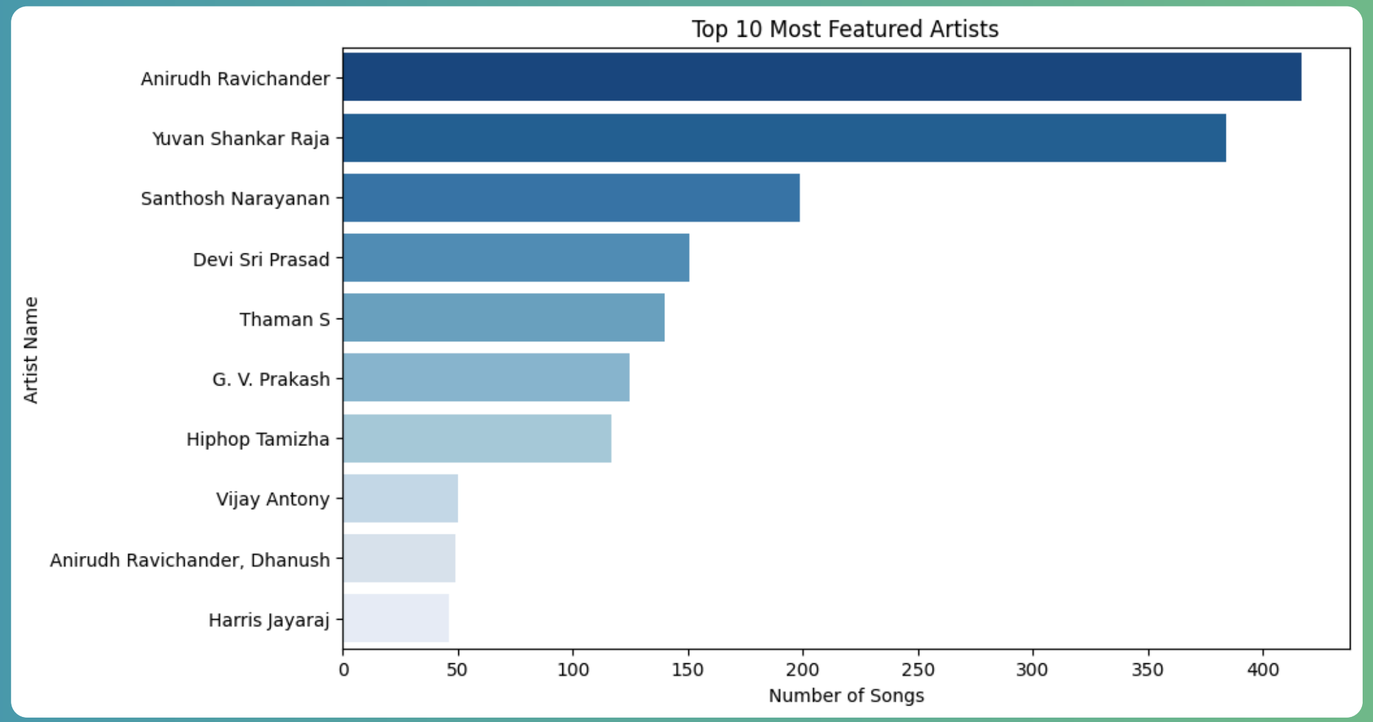
album_name	count	percent_of_dataset
Pudhupettai (Original Motion Picture Soundtrack)	36	0.59
Manmadhan (Original Motion Picture Soundtrack)	32	0.52
Manmadhan (Original Background Score)	32	0.52
Playback: Madras Melodies - Soulful Tamil Melodies	31	0.50
Maan Karate Special (Original Motion Picture Soundtrack)	30	0.49
Whistle Podu	30	0.49
Playback: Justu Kuthu - Best Tamil Folk Songs	29	0.47
Playback: Kaadhal Galatta - Fun Tamil Love Songs	29	0.47
3 (Original Motion Picture Soundtrack)	27	0.44
Playback: Pure Kaadhal - Enchanting Tamil Love Songs	26	0.42
My Playlist: Yuvanshankar Raja	26	0.42
Drishyam 2 (Original Background Score)	25	0.41
Jigarthanda (Original Motion Picture Soundtrack)	25	0.41
Tholi Prema BGM	25	0.41
Vakeel Saab OST	24	0.39
My Playlist: G.V. Prakash Kumar	24	0.39
Kaalai (Original Motion Picture Soundtrack)	23	0.37
Playback: Dance Machi Dance - The Ultimate Tamil Dance Collection	22	0.36
Varisu Original Sound Track	22	0.36
Sony Music Premier League: Ultimate Dance Collection	22	0.36

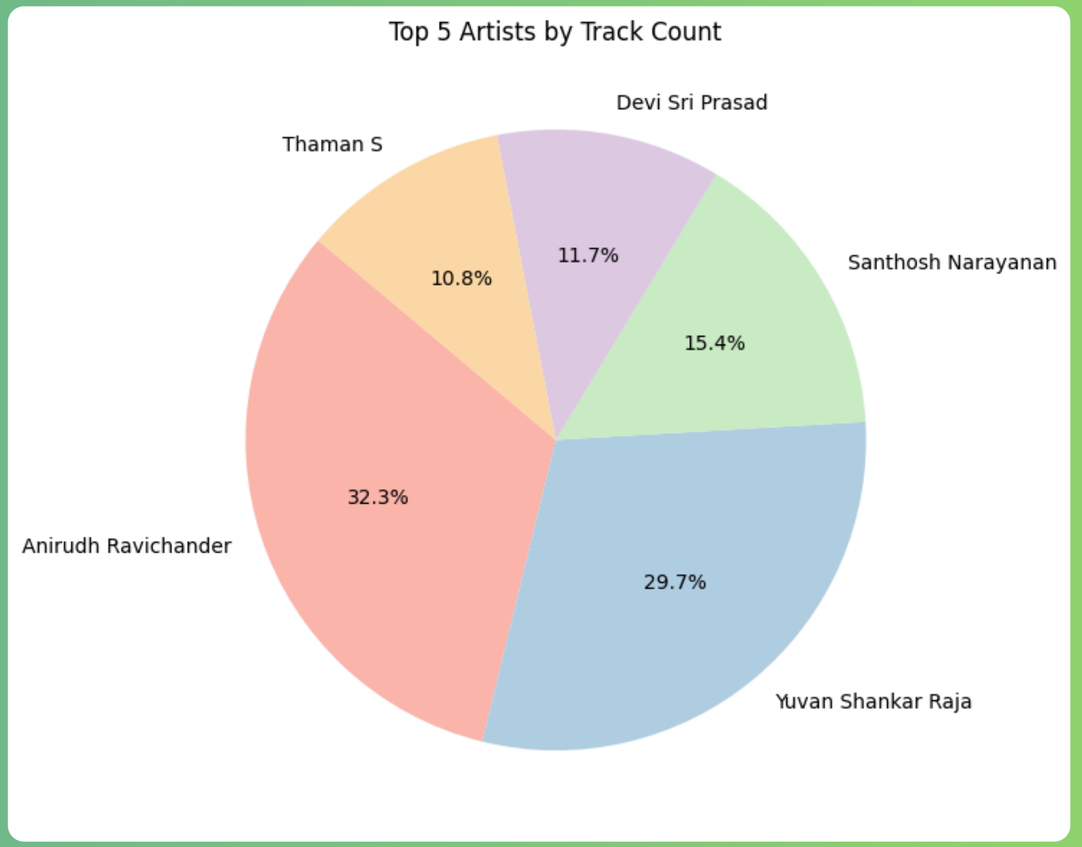
```
df['language'].value_counts()
```

language	count
tamil	4599
unknown	1178
telugu	149
malayalam	101
hindi	93
english	30

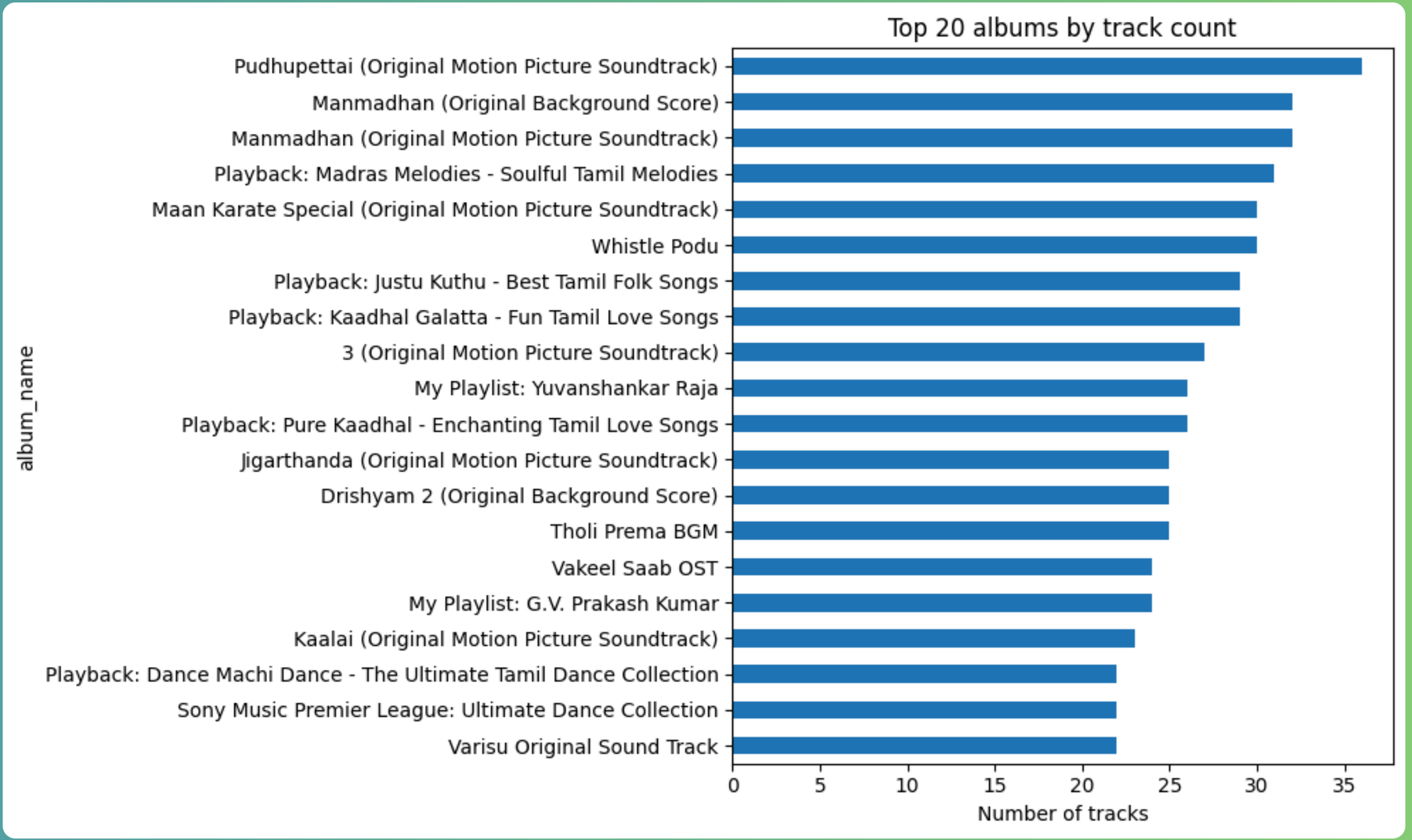
dtype: int64

Distribution of tracks by Artist

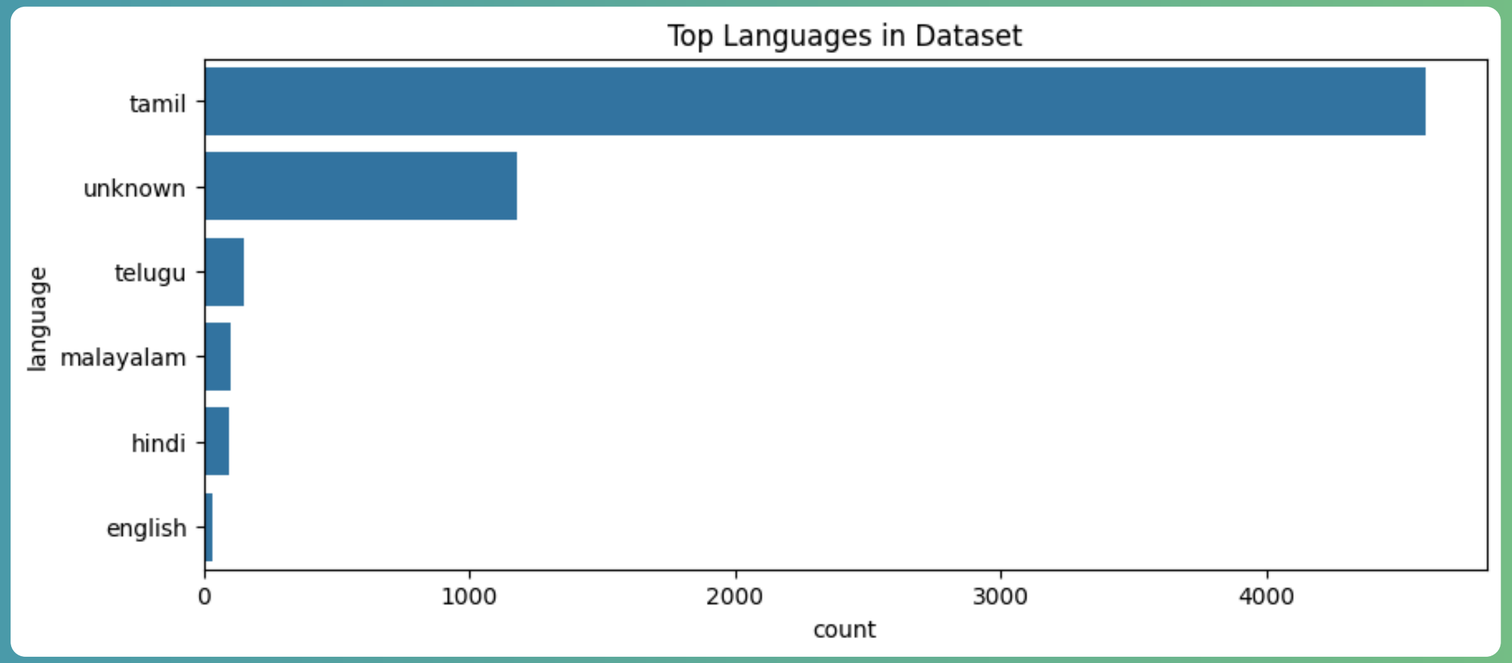


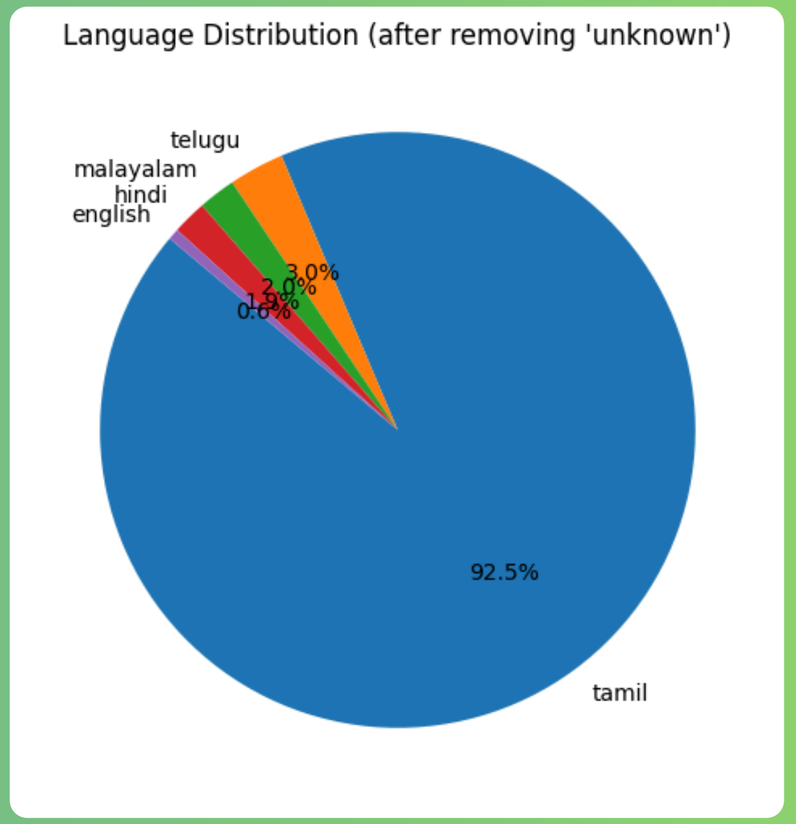


Distribution of tracks by album (top albums)



Distribution of tracks by language





Bivariate Analysis: Feature Relationships

Danceability vs. Mode (Major/Minor)

Violin plots comparing track **Danceability** across major and minor modes indicated a slightly higher mean danceability score for tracks in the Major mode, aligning with a generally more "upbeat" feel.

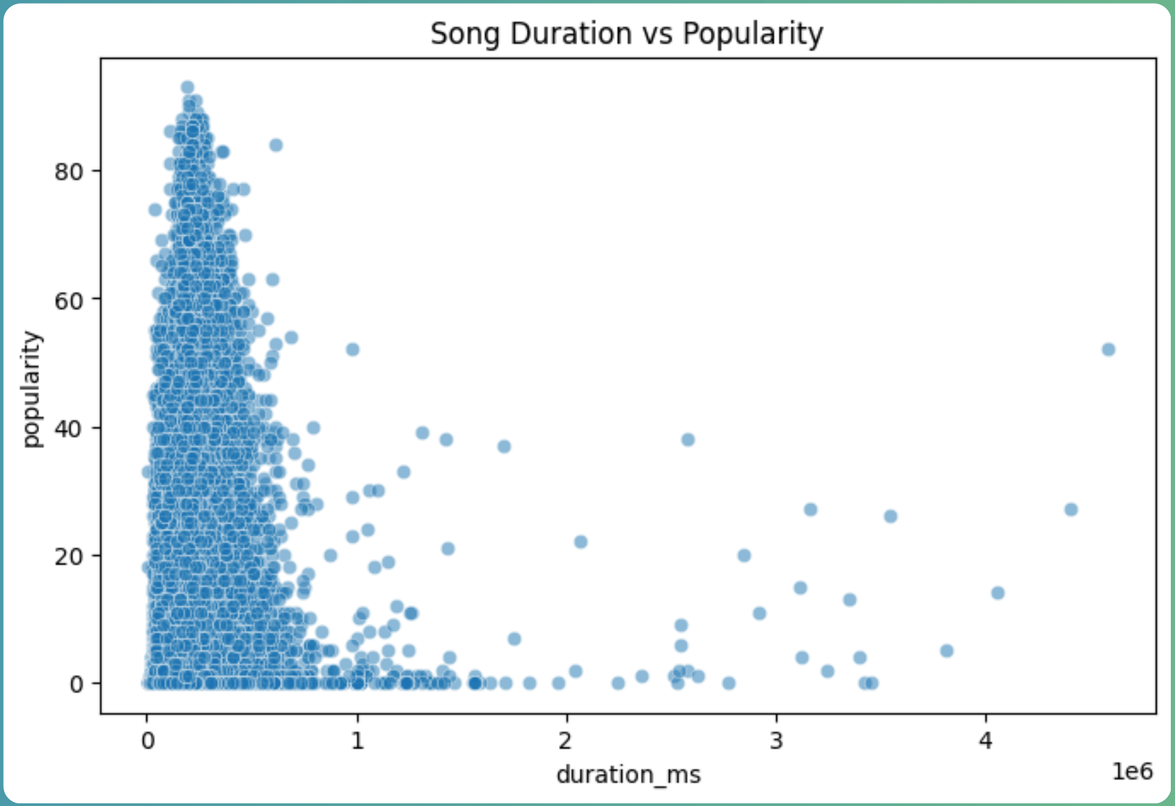
Energy vs. Acousticness

Box plots demonstrated a strong inverse relationship: as expected, tracks with high **Energy** (typically modern, loud) consistently show significantly lower **Acousticness** values (tracks that are not purely acoustic).

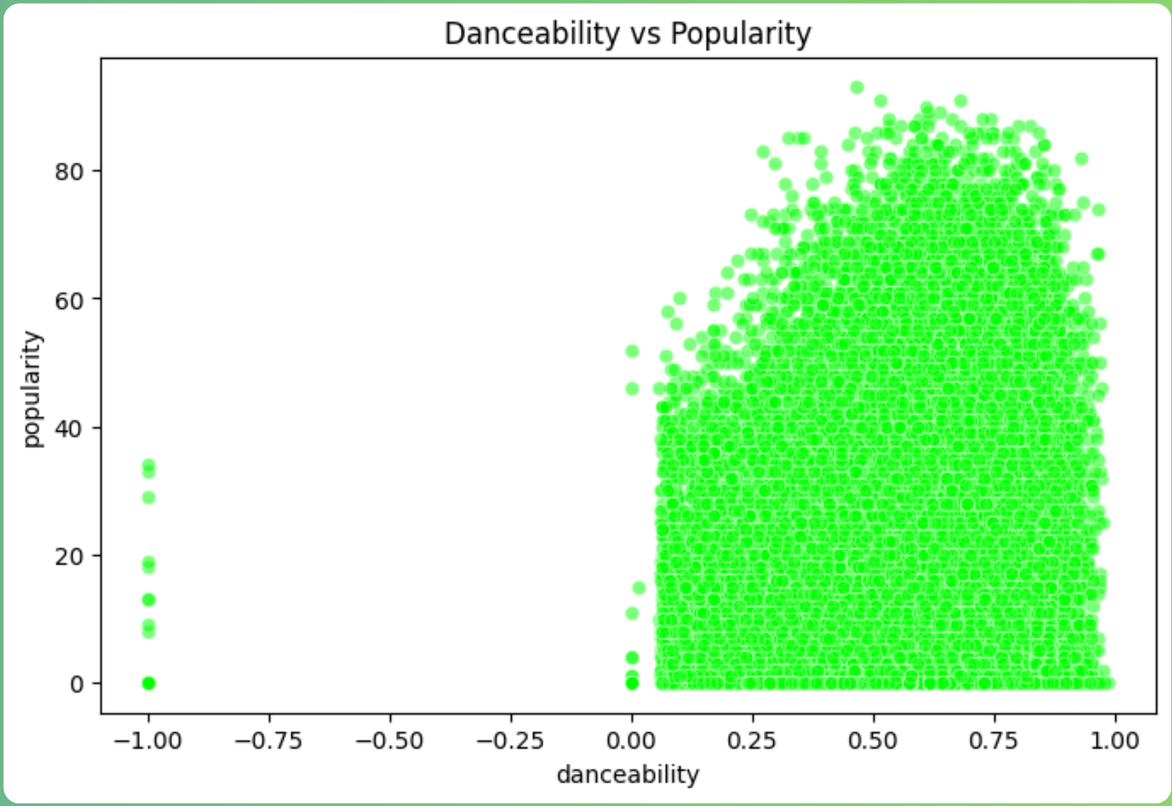
Loudness vs. Popularity

Tracks with maximum **Loudness** ratings tended to correspond with higher mean **Popularity** scores, suggesting that sound engineers may be prioritizing perceived loudness in mainstream production.

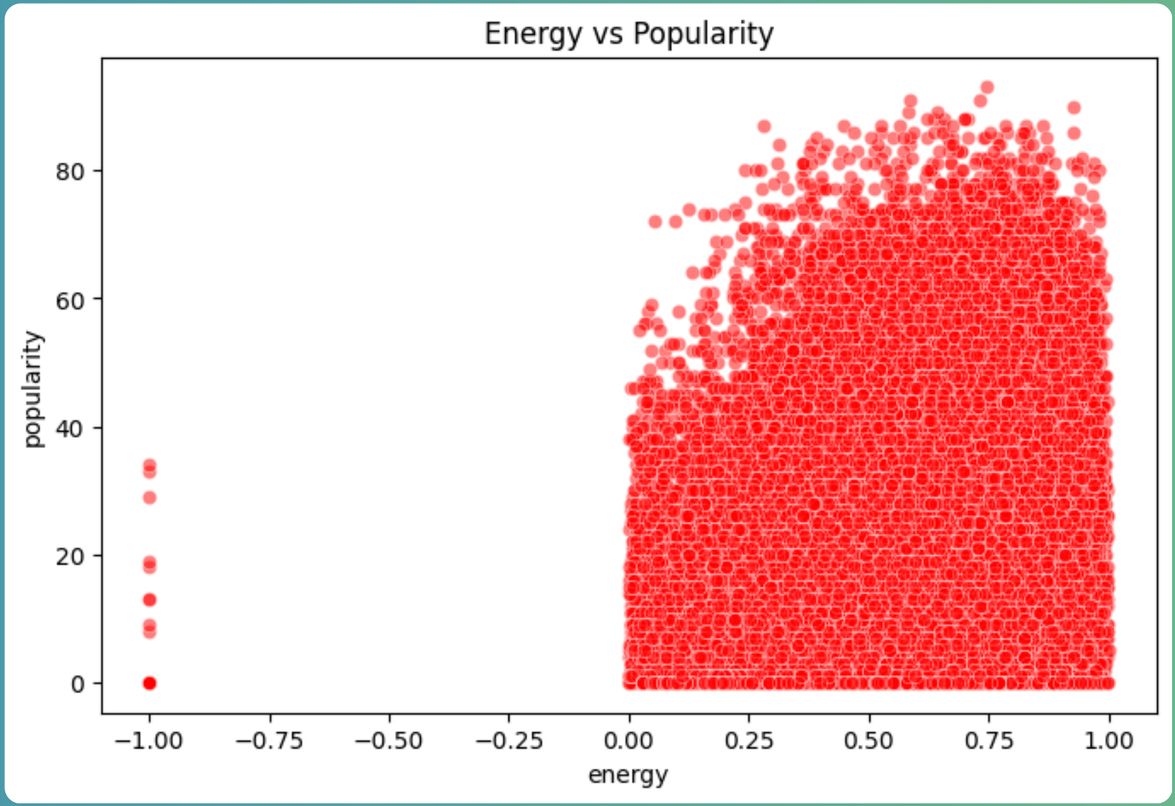
Song Duration vs Popularity



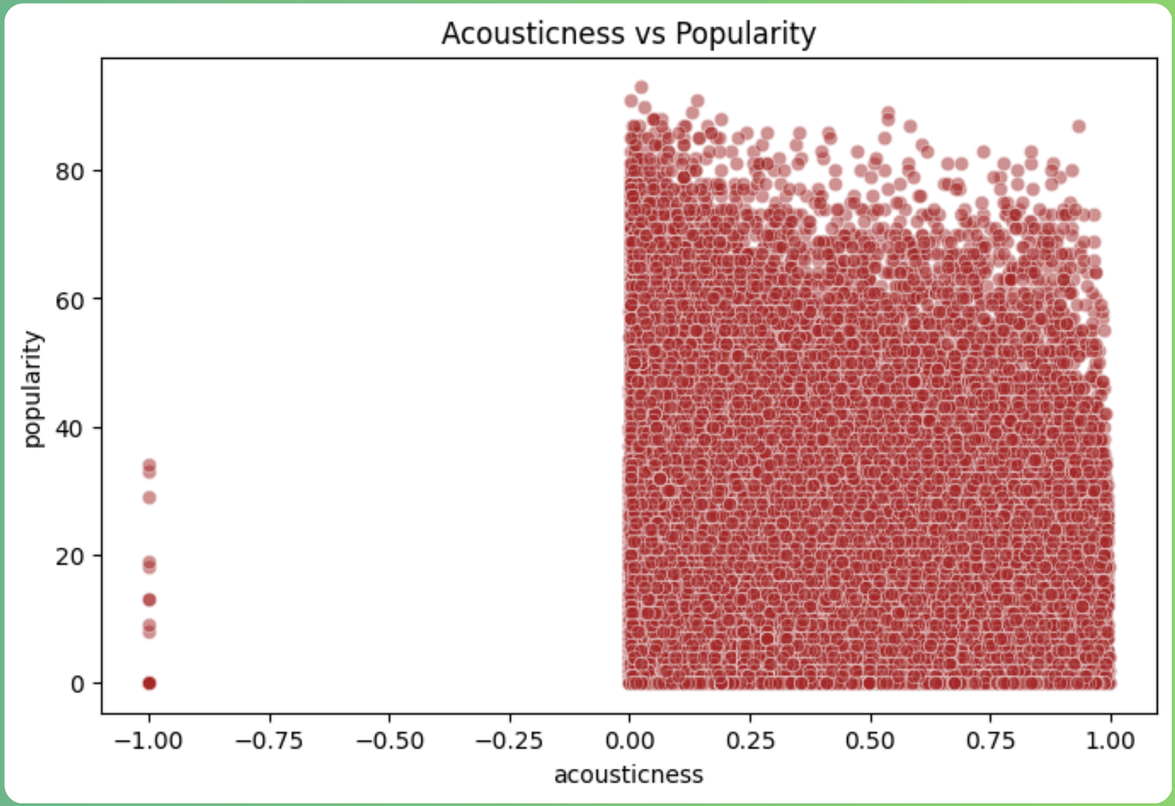
Danceability vs Popularity



Energy vs Popularity



Acousticness vs Popularity

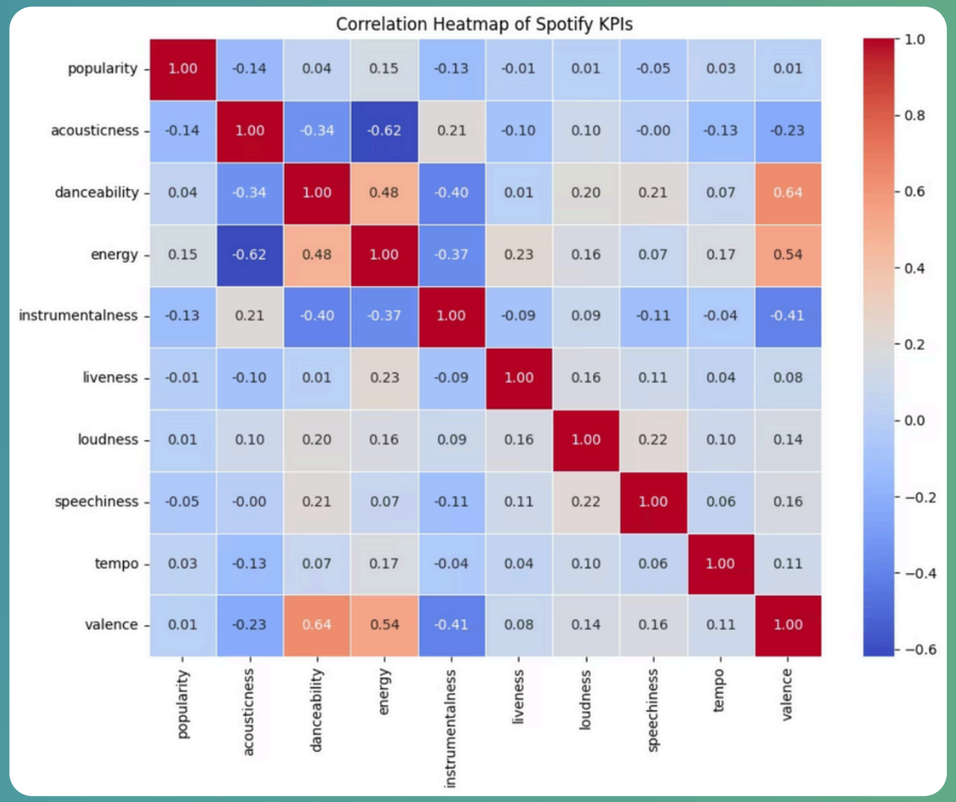


Insights - Bivariate Analysis

- Popularity scores are generally higher for English and Korean songs, while Tamil and Telugu tracks tend to have lower and more consistent popularity.
- Danceability and energy values are elevated for international genres, especially English and Korean, aligning with global pop styles.
- Valence (positivity) remains relatively stable across languages, but Tamil and Hindi tracks show slightly less positive mood overall.
- Tracks in major keys (mode=1) typically have better danceability and slightly greater popularity than those in minor keys.
- Yearly analysis indicates that both energy and danceability have increased over time, mirroring modern production preferences and evolving listener tastes.

Multivariate Analysis: Interacting Forces

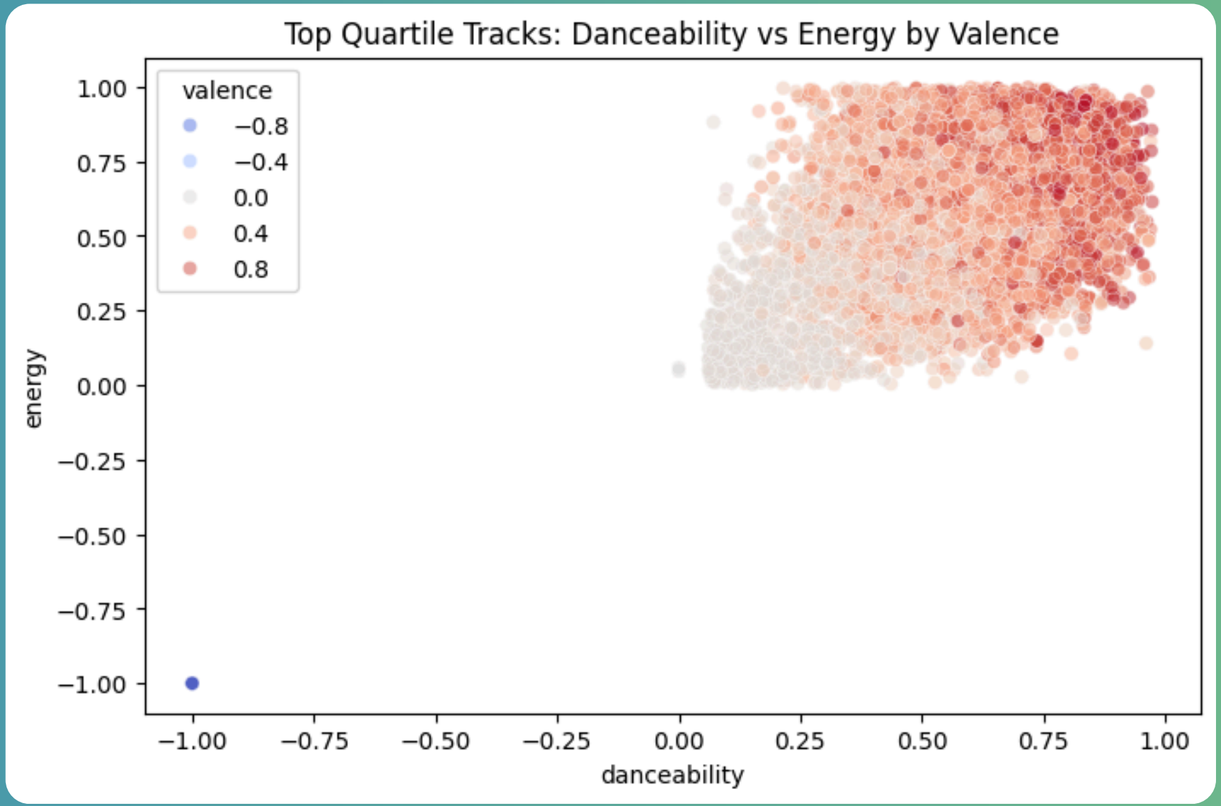
A correlation heatmap provides a comprehensive view of how all numerical acoustic features interrelate, highlighting clusters of highly correlated metrics.



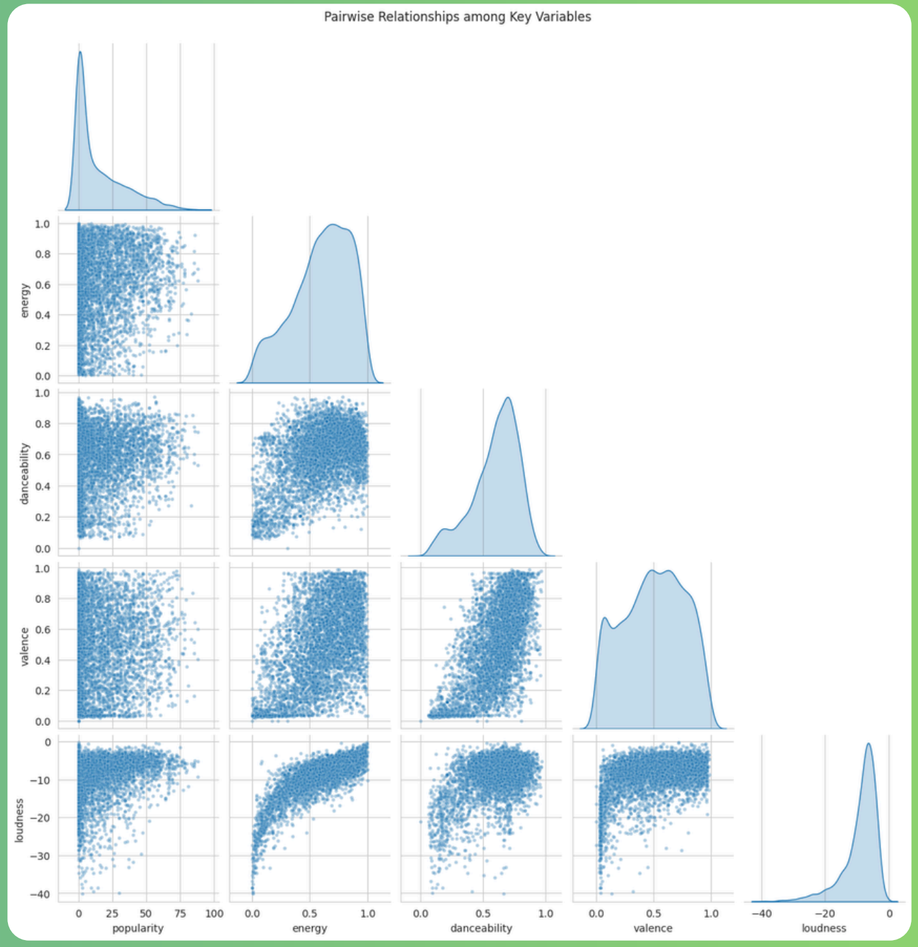
Key correlations observed:

- **Energy and Loudness :** Tracks with higher energy are almost always produced to be louder. This is a common pattern in modern music engineering.
- **Energy and Acousticness :** Highly acoustic tracks have low energy, suggesting a clear separation between raw, live recordings and synthesized, powerful productions.
- **Valence and Danceability :** Tracks perceived as happier (high valence) are also more likely to be highly danceable, indicating a link between positive mood and rhythmic quality.

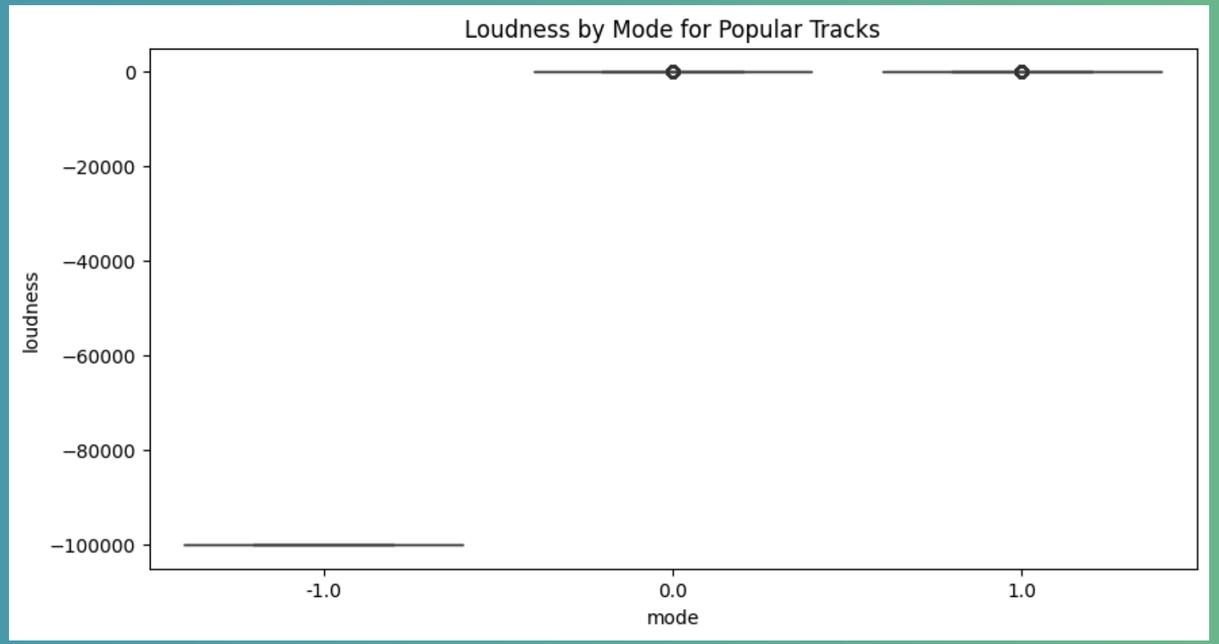
Top Quartile Tracks: Danceability vs Energy by Valence:-



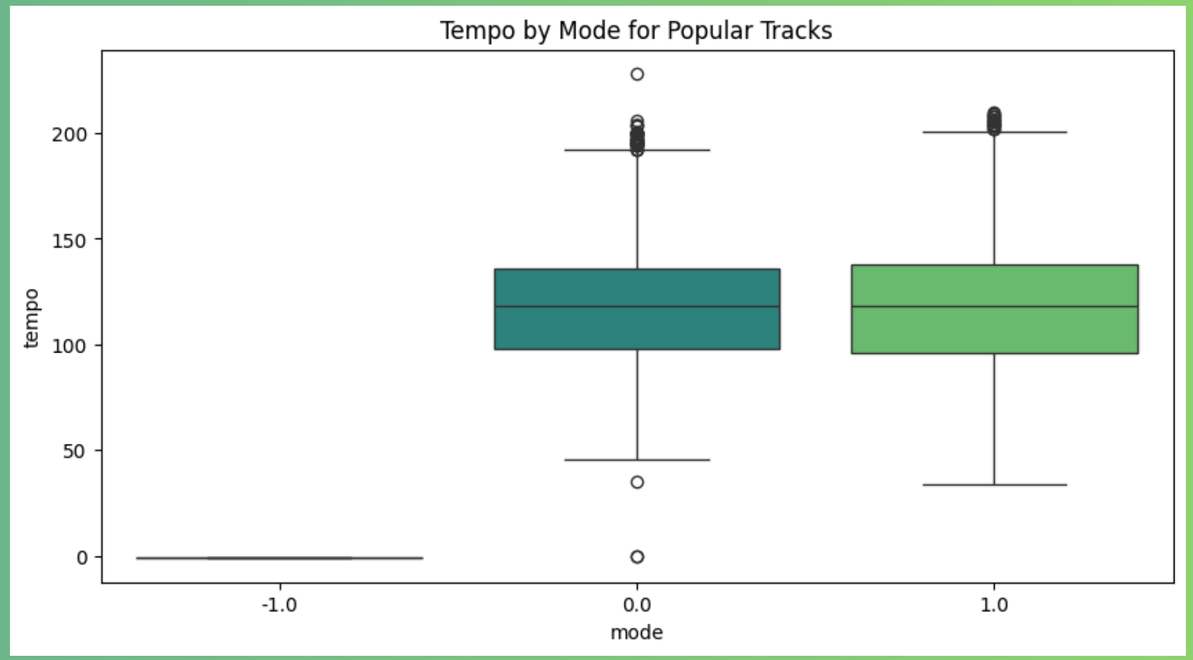
Pairwise Relationships among Key Variables:



Loudness by Mode for Popular Tracks:



Tempo by Mode for Popular Tracks:



Time Series Analysis: The Evolution of Music

1

Popularity Over Years

3

Mean Duration and Liveness Over Years

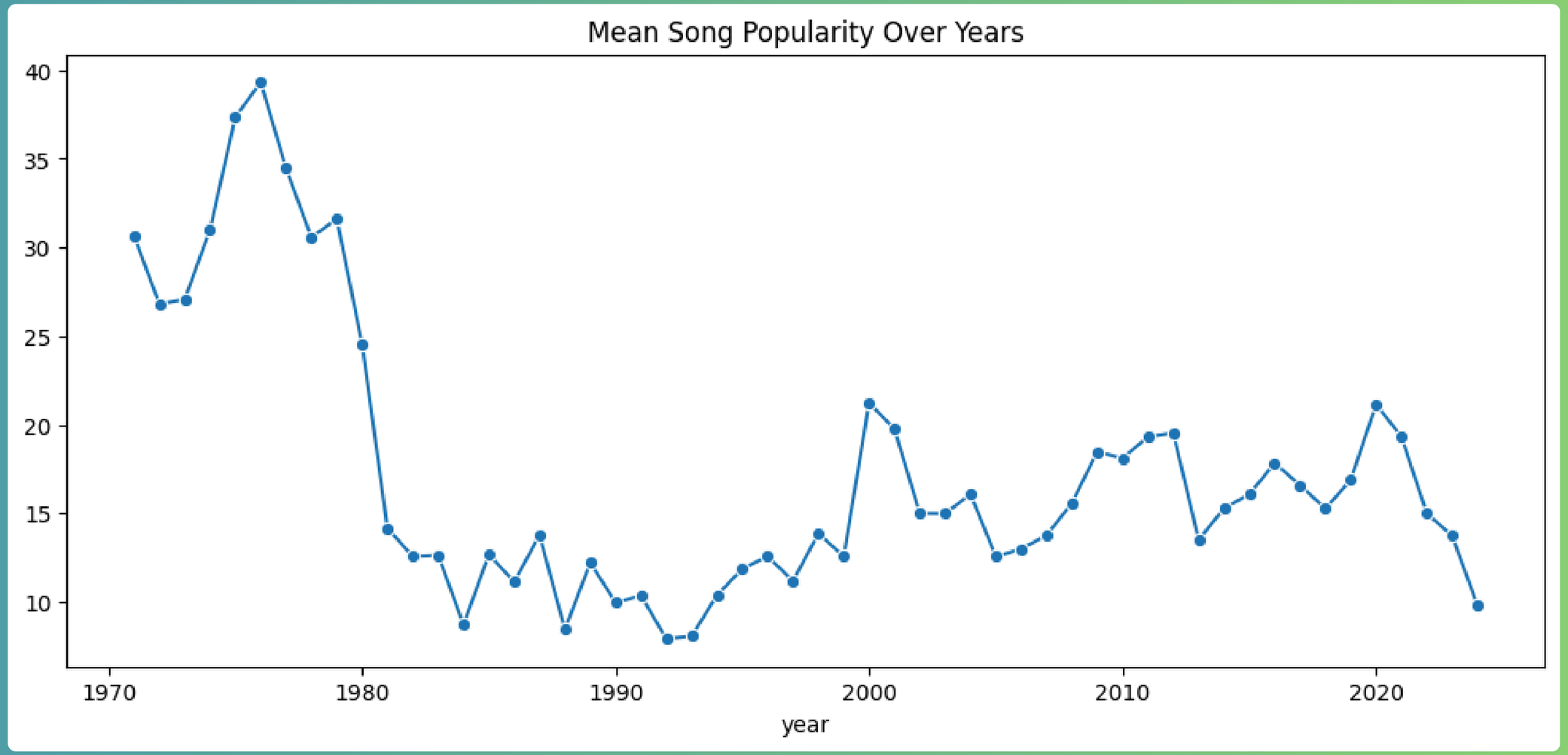
2

Danceability Over Years

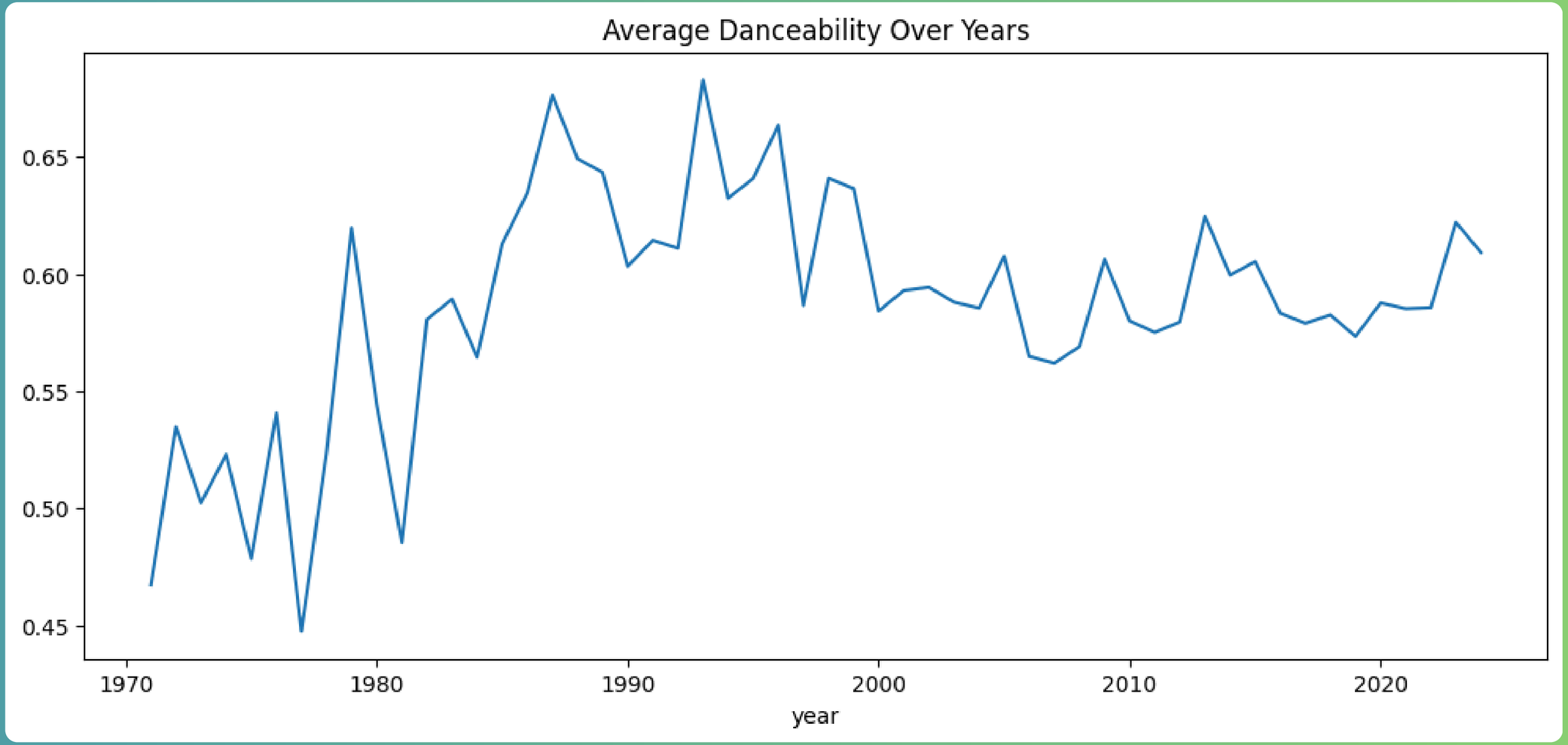
4

Mean Valence & Loudness Over Years

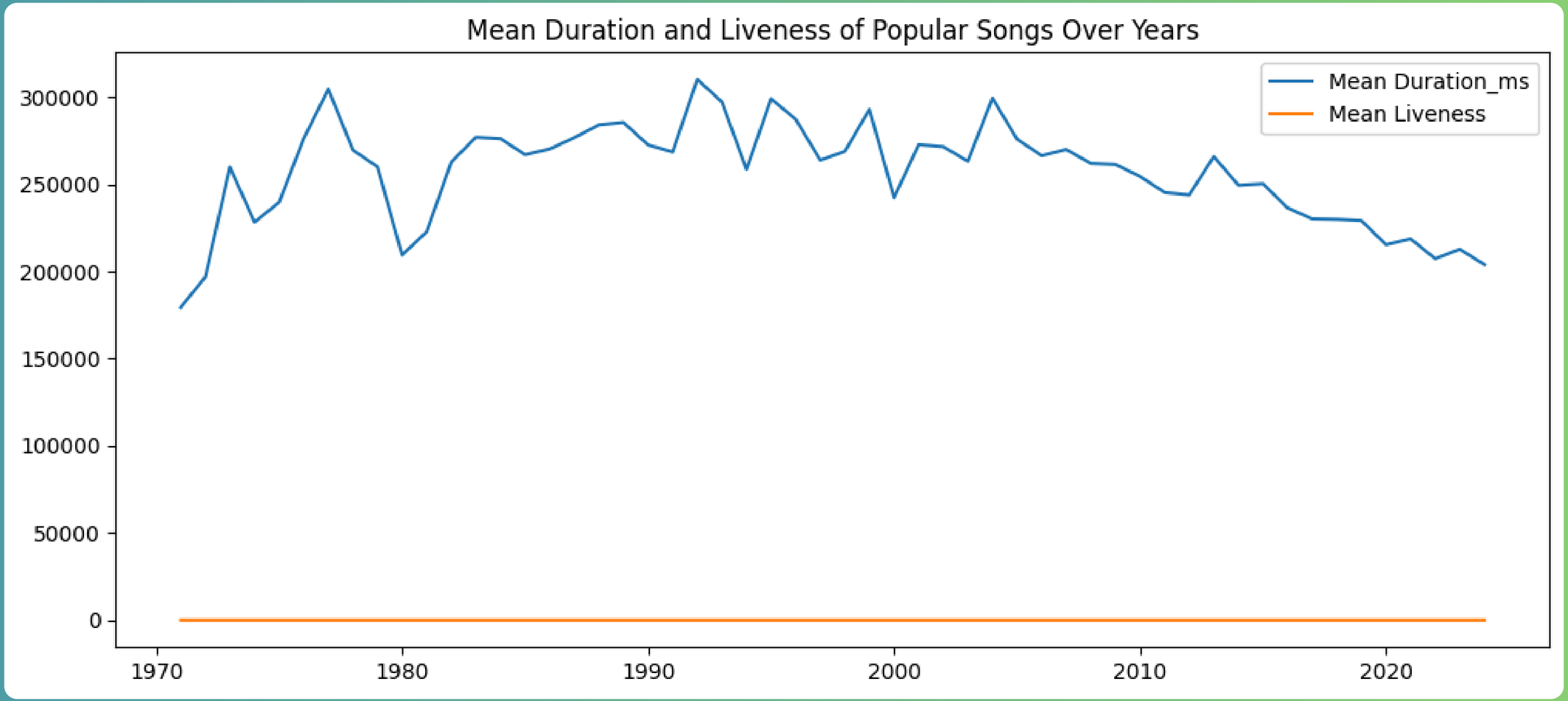
Mean Song Popularity Over Years



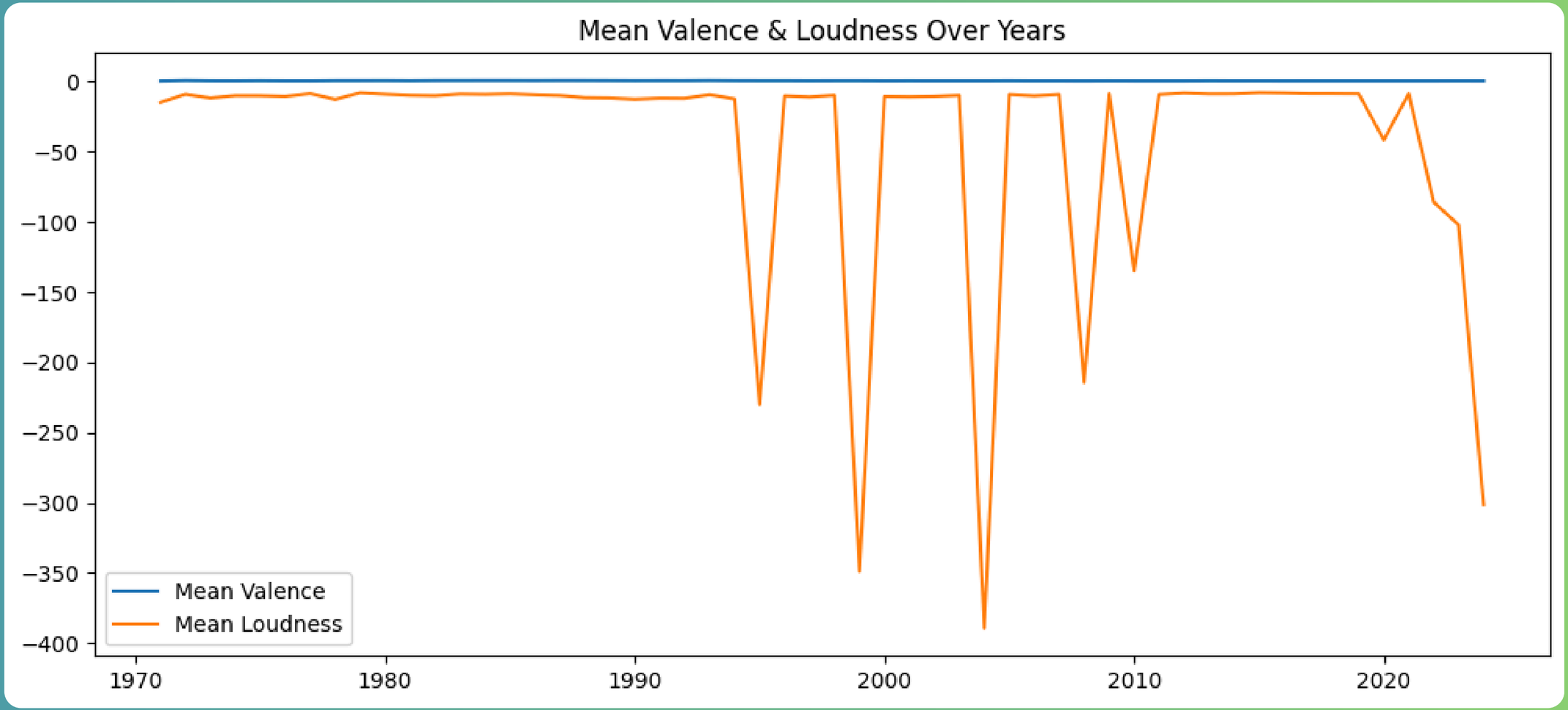
Average Danceability Over Years



Mean Duration and Liveness of Popular Songs Over Years



Mean Valence & Loudness Over Years



Overall Insights in Time Series Analysis

- Average song popularity peaked in the late 1970s, then declined and has stayed relatively flat from the 1990s onward.
- The number of tracks and language diversity in music has rapidly increased since 2010, especially after 2020.
- Mean song duration was highest between 1975 and 2000, then gradually shortened after 2010, while mean liveness remained nearly unchanged.
- Average danceability has shown a slow upward trend since the 1980s, indicating songs are generally becoming more danceable.
- Mean valence (happiness) remained stable, while mean loudness has significant negative outliers in recent years, likely due to data issues or changes in production techniques.

Outlier Detection: Identifying Anomalous Tracks

Using the Interquartile Range (IQR) technique, we isolated tracks whose feature values fall significantly outside the norm. These outliers often represent unique content or recording anomalies.

Extremely High Duration

Outliers here are predominantly classical compositions or long-form spoken word content, differing dramatically from the average 3-4 minute pop track. These tracks require separate categorization for accurate consumption modeling.

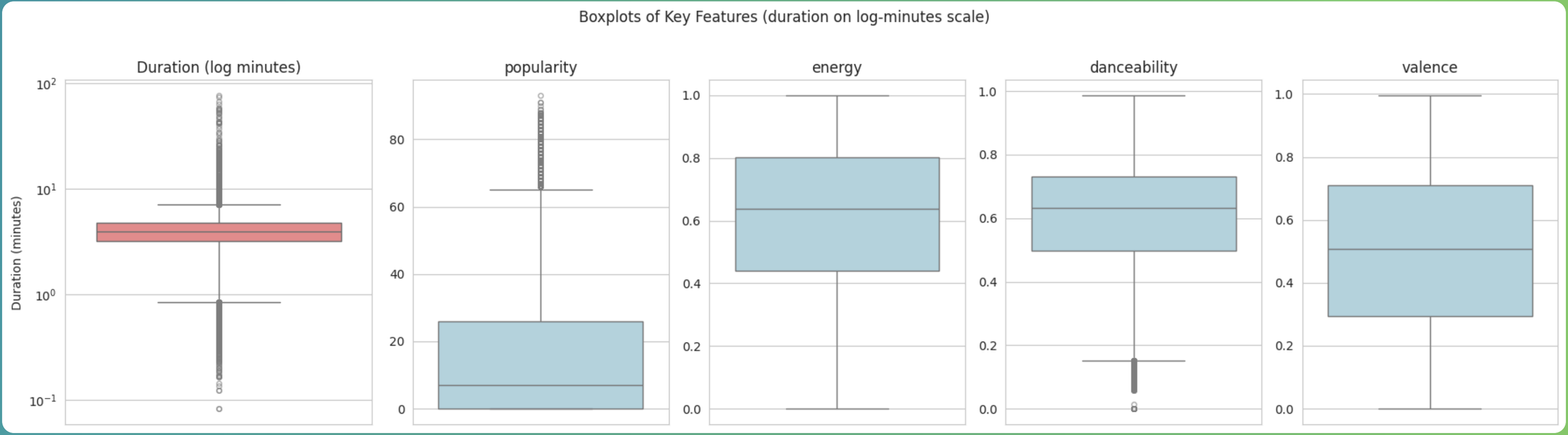
Low Energy, High Popularity

Tracks that defy the Loudness/Energy trend, such as ambient soundscapes or highly specific niche genres, often maintain high popularity among dedicated user segments. This highlights the importance of non-mainstream content.

Anomaly in Tempo

Tracks with excessively high (e.g., >200 BPM) or low (e.g., <50 BPM) tempos are flagged. These often represent production errors or highly experimental tracks, which can impact recommendation engine tuning.

Boxplots of Key Features (duration on log-minutes scale)



Insights — Outlier Analysis

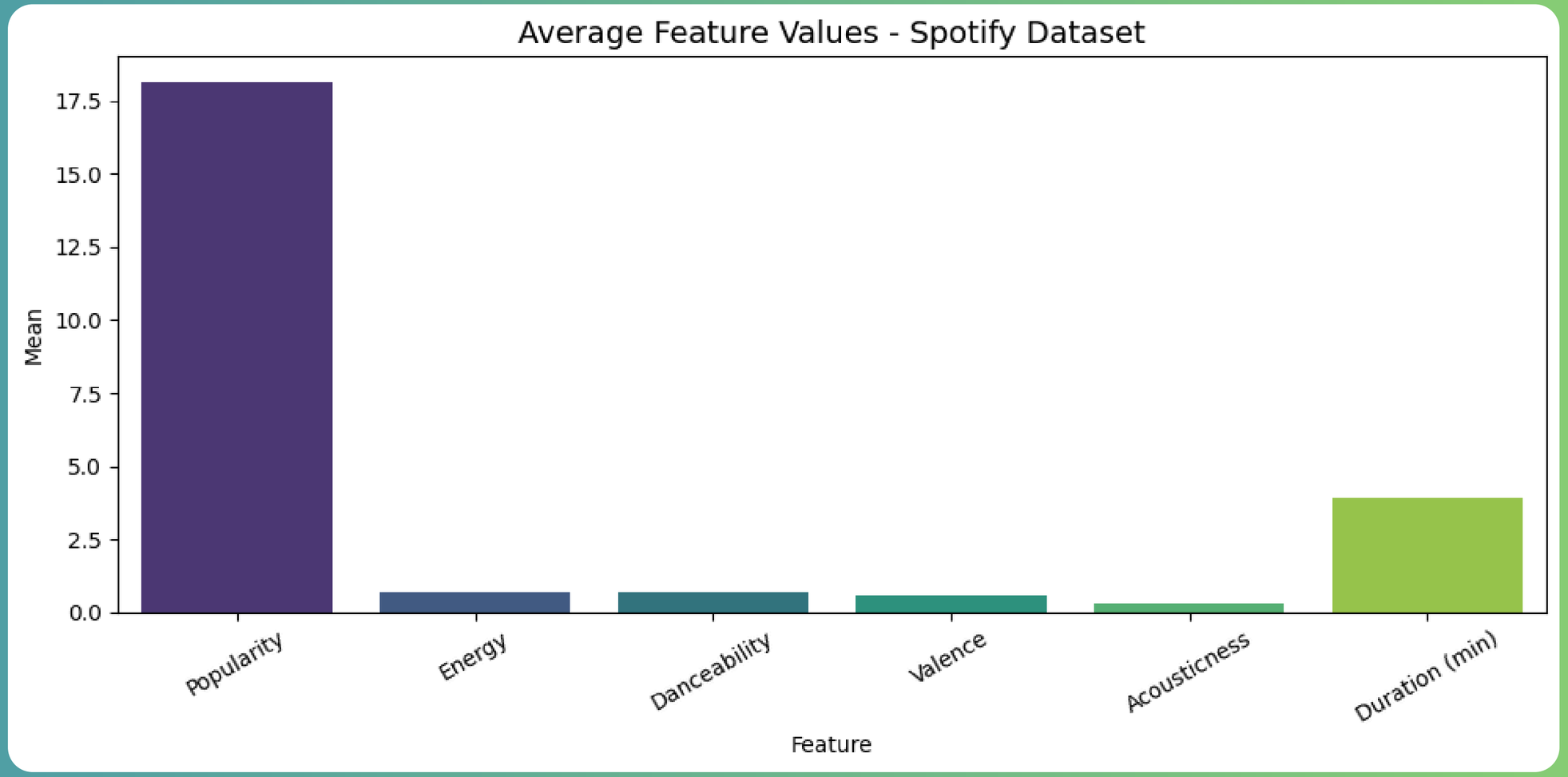
- Extreme values were observed in duration, loudness, and popularity, showing the presence of genuine outliers in the dataset.

- Very long durations often indicate podcasts, live recordings, or extended mixes, not standard songs.
- Loudness outliers can reveal tracks that are either poorly normalized/acoustic (very quiet) or mastered aggressively (very loud).

- Only a small number of tracks reach popularity scores above 80, confirming that very high audience engagement is rare.

- Energy and danceability display consistency, with few outlier values compared to other features.
- Outliers are best handled based on context—remove only those that misrepresent general song trends, as some may highlight special or interesting cases in music data.

Average Feature Values - Spotify Dataset



Final Insights and Strategic Recommendations

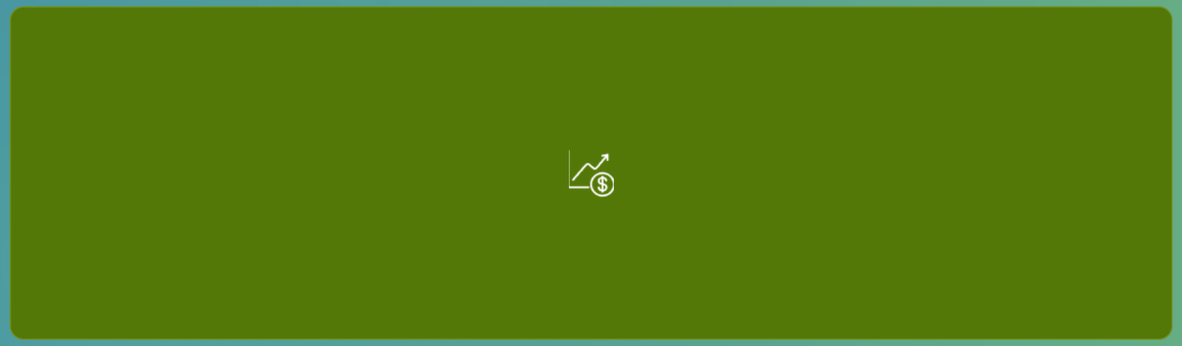
Synthesizing the analysis across all features yields critical strategic direction for content curation and algorithm tuning.

Recommendation 3: Optimize for Modern Trends

Prioritize the creation and promotion of playlists that align with the observed upward trends in **Energy**, **Danceability**, and **Loudness** to capture mainstream listener engagement.

Recommendation 2: Refine Niche Discovery

Develop separate recommendation models tailored specifically for outlier content (long duration, very low energy) to serve niche communities effectively, increasing long-tail content consumption.



Recommendation 1: Validate Correlation Filters

Leverage the strong positive correlation between **Valence** and **Danceability** to build highly effective "mood" or "party" recommendation filters, improving user experience and session length.

THANK YOU!